

# Przetwarzanie języka naturalnego i Text Mining

Dr Bogna Zaczny

Katedra Inżynierii Wiedzy

Uniwersytet Ekonomiczny w Katowicach

# Agenda

- Analiza nastroju
- Analiza wątków

**ANALIZA NASTROJU**

# Analiza nastroju

Sentiment analysis to kierunek badań analizujący ludzkie:

opinie, uczucia, oceny, ekspertyzy, postawy, poglądy, emocje,  
wobec podmiotów, takich jak produkty, usługi, organizacje, osoby fizyczne, sprawy, wydarzenia, tematy i ich atrybuty.

# Analiza nastroju

Ze względu na różne podejścia można wyróżnić:

- sentiment analysis,
- opinion mining,
- opinion extraction,
- sentiment mining,
- subjectivity analysis,
- affect analysis,
- emotion analysis,
- review mining.

# Analiza nastroju

Identyfikacja ładunku emocjonalnego wypowiedzi i zaszeregowaniu go do jednej ze wskazanych kategorii wypowiedzi.

# Analiza nastroju - podejścia

- Metody słownikowe:
  - Budowane ręcznie / (pół)automatycznie
- Metody statystyczne:
  - Zbiory trenujące
  - Różne rodzaje zmiennych opisujących (*features*):
    - Słowa
    - Współwystępowanie słów
    - Interpunkcja
    - Składnia
    - Emoticony

<https://yougov.co.uk/news/2018/10/02/how-good-good/>

# Metody słownikowe

Słowniki:

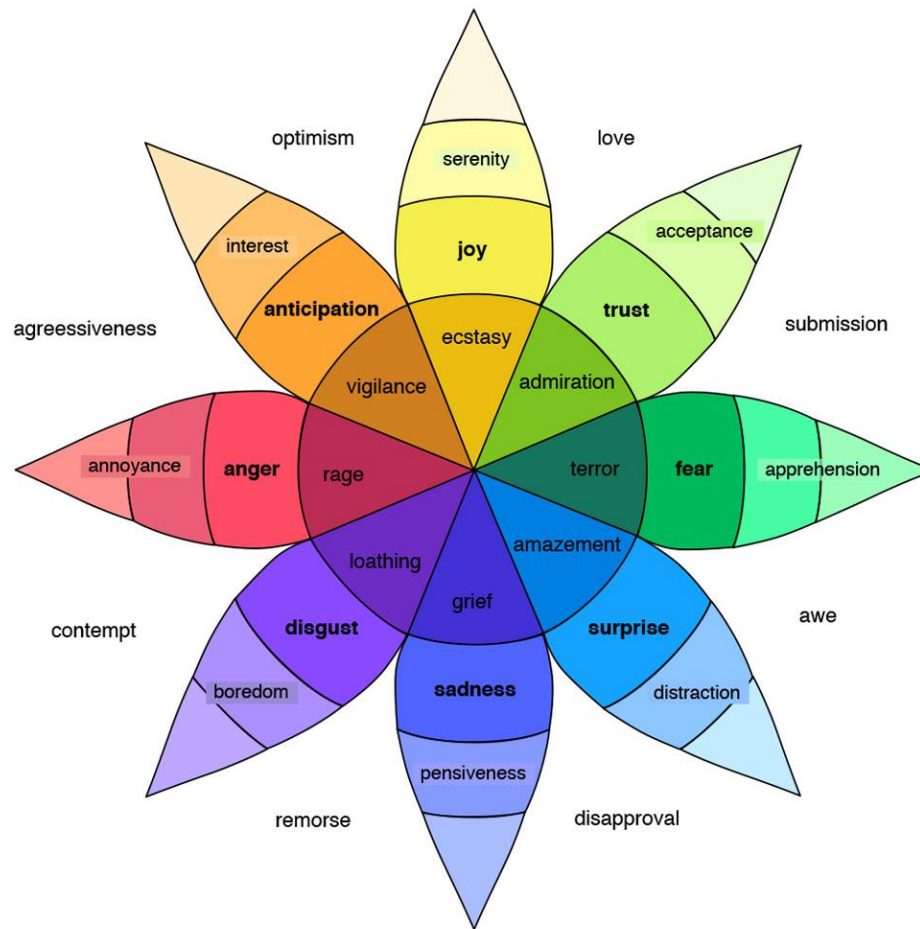
- Finn Årup Nielsen
- Bing Liu and collaborators
- Saif Mohammad and Peter Turney



# Metody słownikowe

<b>Dok</b>	<b>Tekst</b>	<b>syuzhet</b>	<b>bing</b>	<b>afinn</b>	<b>nrc</b>
<b>1</b>	Bought it as a random gift for my daughter and she loves it!!!!	2.10	1	2	2
<b>2</b>	I bought this for my daughter and family. Everybody loves Alexa!	1.60	1	0	1
<b>3</b>	Pretty disappointing. It's pretty worthless.	-1.25	-1	-3	-1
<b>4</b>	Disappointing and frustrating ? I'm returning Echo	-1.50	-2	-4	-1

# Plutchik's wheel of emotions



# Plutchik's wheel of emotions

Dok	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	negative	positive
1	0	1	0	0	2	0	1	0	0	2
2	0	0	0	0	1	0	0	0	0	1
3	1	1	1	0	1	2	0	1	2	1
4	0	0	0	0	0	1	0	0	1	0

[1] "Bought it as a random gift for my daughter and she loves it!!!!!"

[2] "I bought this for my daughter and family. Everybody loves Alexa!"

[3] "Pretty disappointing. It's pretty worthless."

[4] "Disappointing and frustrating ? I'm returning Echo"

# **ANALIZA WĄTKÓW**

# Topic modeling

- Modele tematyczne pozwalają na organizowanie i streszczanie dużych zbiorów danych

# Wykrywanie tematu(ów) dokumentów

- tf-idf – dla wybranego zbioru termów i każdego dokumentu liczony jest współczynnik tf-idf
- Analiza ukrytych grup semantycznych (*Latent Semantic Analysis*, **LSA**) – wykorzystanie dekompozycji SVD do wykrycia podprzestrzeni cech tf-idf
- Modele probabilistyczne:
  - unigramy (*unigrams*)
  - mieszanina unigramów (*mixture of unigrams*)
  - probabilistyczna analiza ukrytej semantyki (*Probabilistic Latent Semantic Analysis*, **PLSA**)
  - alokacja ukrytej zmiennej Dirichleta (*Latent Dirichlet Allocation*, **LDA**)
  - hierarchiczny proces Dirichleta (*Hierarchical Dirichlet Process*)

# Analiza ukrytych grup semantycznych

# Analiza ukrytych grup semantycznych

## **Rozkład według wartości osobliwych**

(*SVD Singular Value Decomposition*) to pewien rozkład (dekompozycja) macierzy na iloczyn trzech specyficznych macierzy.

Metoda matematyczna stosowana m.in. w analizie statystycznej służąca do redukcji wymiaru macierzy.



# Analiza ukrytych grup semantycznych

	D1	D2	D3	D4
linux	3	4	1	0
modem	4	3	0	1
the	3	4	4	3
clutch	0	1	4	3
steering	2	0	3	3
petrol	0	1	3	4

# Analiza ukrytych grup semantycznych

	D1	D2	D3	D4
linux	3	4	1	0
modem	4	3	0	1
the	3	4	4	3
clutch	0	1	4	3
steering	2	0	3	3
petrol	0	1	3	4

# Analiza ukrytych grup semantycznych

6x4	d1	d2	d3	d4
t1				
t2				
t3				
t4				
t5				
t6				

# Analiza ukrytych grup semantycznych

6x4	d1	d2	d3	d4
t1				
t2				
t3				
t4				
t5				
t6				

=

6x4	w1	w2	w3	w4
t1				
t2				
t3				
t4				
t5				
t6				

X

<b>waż</b>			
	<b>ność</b>		
		<b>wąt</b>	
			<b>ku</b>

X

4x4	d1	d2	d3	d4
w1				
w2				
w3				
w4				

# Analiza ukrytych grup semantycznych

3	4	1	0
4	3	0	1
3	4	4	3
0	1	4	3
2	0	3	3
0	1	3	4

 $=$ 

	To1	To2	To3	To4
Te1	-0.33	-0.53	0.37	-0.14
Te2	-0.32	-0.54	-0.49	0.35
Te3	-0.62	-0.10	0.26	-0.14
Te4	-0.38	0.42	0.30	-0.24
Te5	-0.36	0.25	-0.68	-0.47
Te6	-0.37	0.42	0.02	0.75

 $\times$ 

Topic Importance

11.4
6.27
2.22
1.28

 $\times$ 

	D1	D2	D3	D4
To1	-0.42	-0.48	-0.57	-0.51
To2	-0.56	-0.52	0.45	0.46
To3	-0.65	0.62	0.28	-0.35
To4	-0.30	0.34	-0.63	0.63

# Analiza ukrytych grup semantycznych

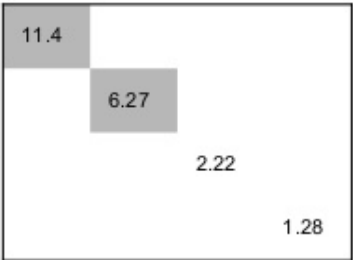
3	4	1	0
4	3	0	1
3	4	4	3
0	1	4	3
2	0	3	3
0	1	3	4

**=**

	To1	To2	To3	To4
Te1	-0.33	-0.53	0.37	-0.14
Te2	-0.32	-0.54	-0.49	0.35
Te3	-0.62	-0.10	0.26	-0.14
Te4	-0.38	0.42	0.30	-0.24
Te5	-0.36	0.25	-0.68	-0.47
Te6	-0.37	0.42	0.02	0.75

**X**

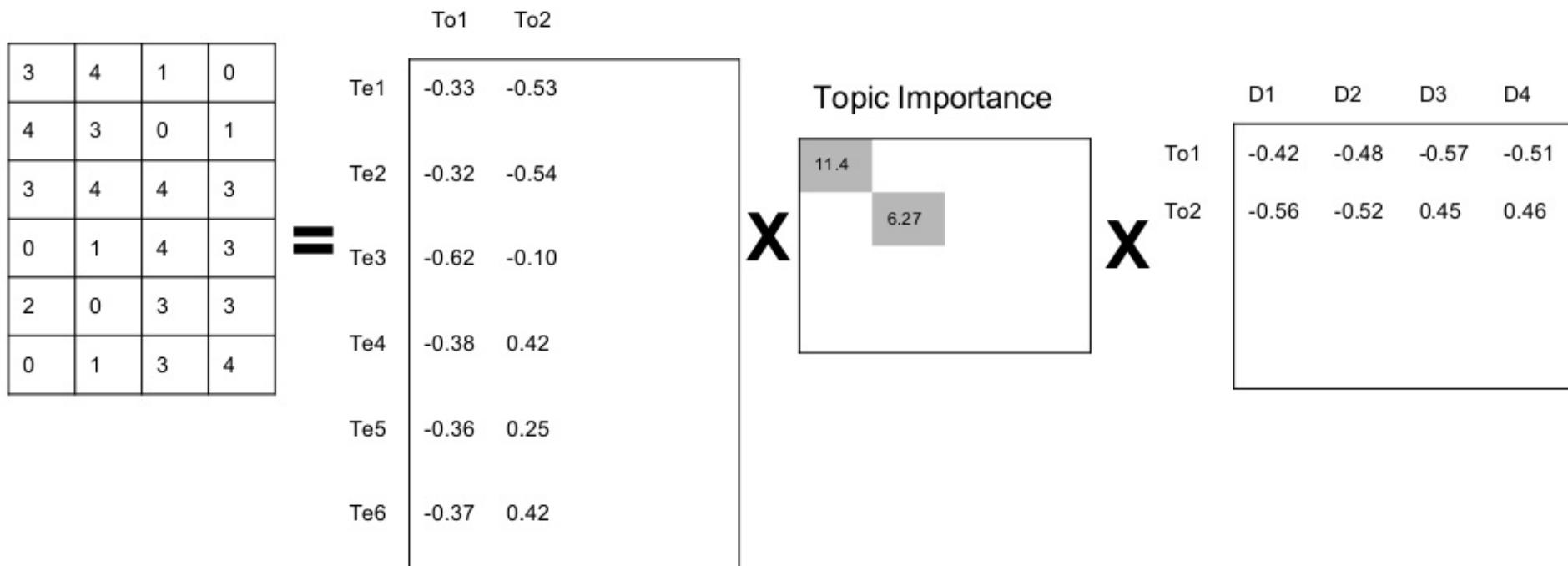
Topic Importance



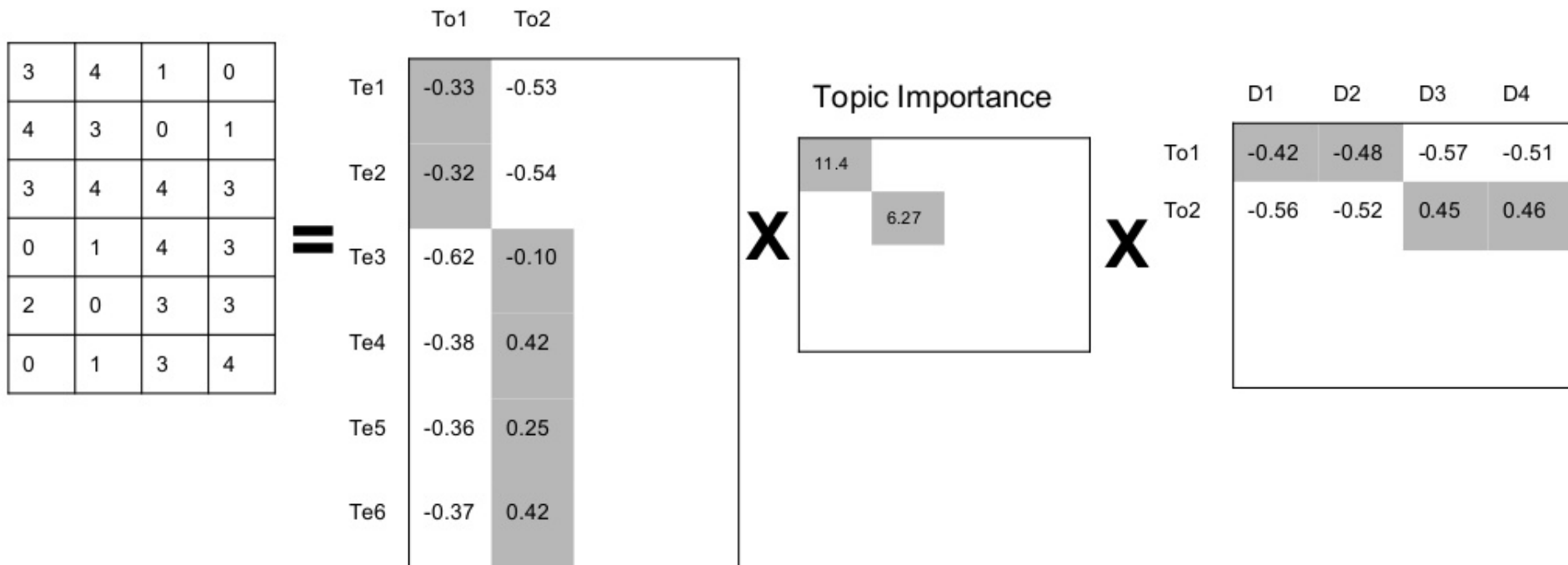
**X**

	D1	D2	D3	D4
To1	-0.42	-0.48	-0.57	-0.51
To2	-0.56	-0.52	0.45	0.46
To3	-0.65	0.62	0.28	-0.35
To4	-0.30	0.34	-0.63	0.63

# Analiza ukrytych grup semantycznych



# Analiza ukrytych grup semantycznych





# Analiza ukrytych grup semantycznych

**Word assignment to topics**

					IT	cars	
3	4	1	0	= =	linux	-0.33	-0.53
4	3	0	1		modem	-0.32	-0.54
3	4	4	3		the	-0.62	-0.10
0	1	4	3		clutch	-0.38	0.42
2	0	3	3		steering	-0.36	0.25
0	1	3	4		petrol	-0.37	0.42

**Topic Importance**

11.4	
	6.27

**X** IT  
cars

**Topic distribution across documents**

	D1	D2	D3	D4
IT	-0.42	-0.48	-0.57	-0.51
cars	-0.56	-0.52	0.45	0.46

# Podejście probabilistyczne

# Podejście probabilistyczne

- Zamiast dokonywać redukcji wymiarów, poszukuje się „mieszanki” rozkładów (słów w wątku oraz wątków w dokumencie), które są najbardziej prawdopodobne w dokumencie.
- Definiowany jest model statystyczny tworzenia (generowania) dokumentu. Co określane jest procesem generującym.

# Proces generujący

Statystyczny model generujący - opisuje (hipotetyczny) proces losowy, generujący kolejne słowa każdego dokumentu zgodnie z rozkładem wątków w dokumencie.

W praktyce dany jest tylko tekst, a wątki i ich rozkłady w dokumentach należy estymować.

# Proces generujący - przykład

Artykuł:

- dokument – 50 słów;
- 4 możliwych wątków;
- 6 słów.

W jaki sposób może powstać artykuł?

# Proces generujący - przykład

Artykuł:

- dokument – 50 słów;
- 4 możliwych wątków;
- 6 słów.

W jaki sposób może powstać artykuł?

- 1) 1. słowo artykułu: rzucamy kostką do gry, wynik mówi nam, z którego wątku będzie pobrane słowo. Załóżmy, że jest to wątek **1** (*IT*);
- 2) Następnie rzucamy kolejną kostką, aby wybrać słowo, które opisuje wątek **1**. Załóżmy, że jest to słowo **1** (*Linux*) ;
- 3) Proces ten jest powtarzany dla wszystkich 50 słów w dokumencie.

# Proces generujący - przykład

Istotne założenie: Kości są ważone!!!

- Pierwsza kość wybierająca wątek kładzie większy ciężar na wątek IT, niż na pozostałe 3 wątki.
- Kość dla wyboru słowa z wątku (IT), kładzie większy ciężar na słowa 'linux' i 'modem'.
- Podobnie kości do wątku 2 (samochody) kładzie większy nacisk na słowo "benzyna" i „sprzęgło"

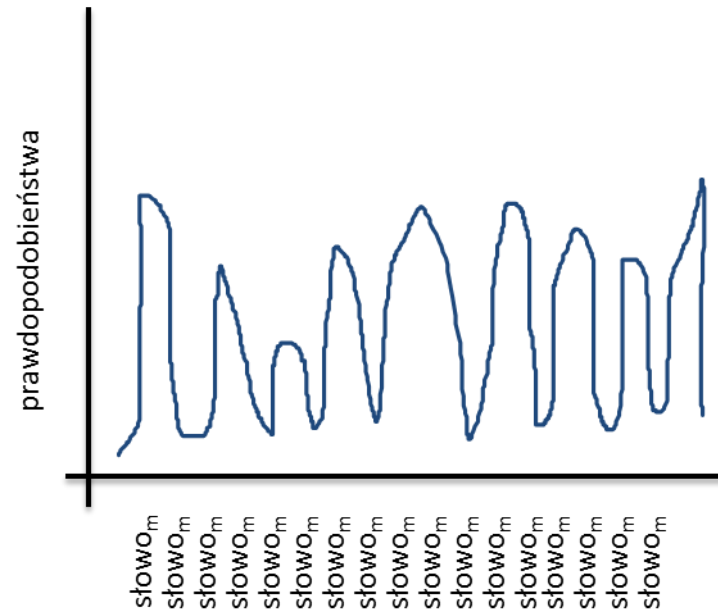
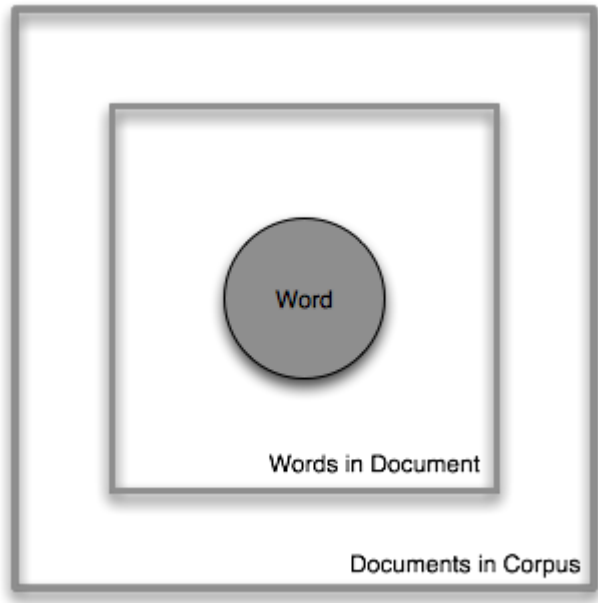
# Proces generujący - demo



# Modele probabilistyczne

# Unigram

# Unigram



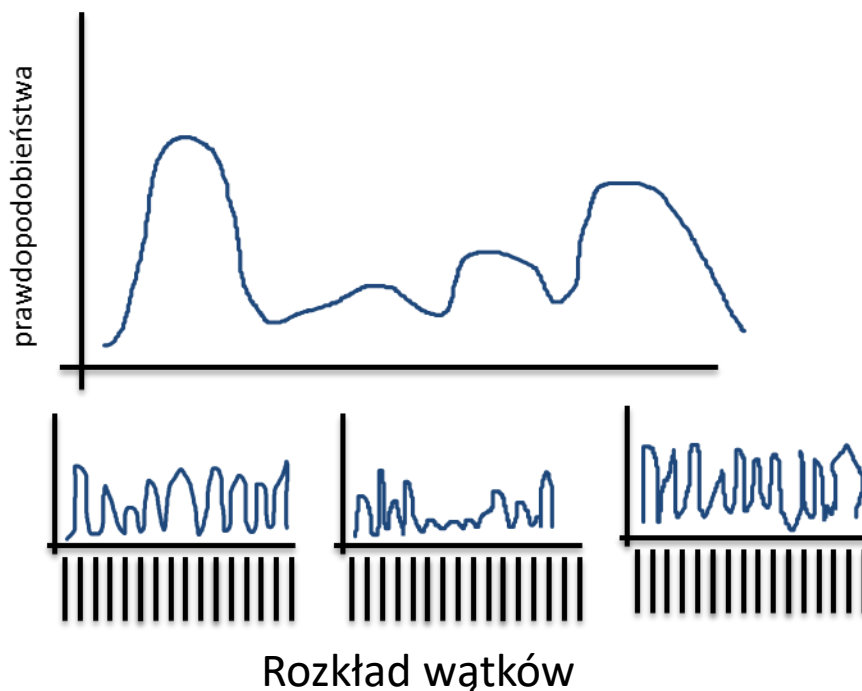
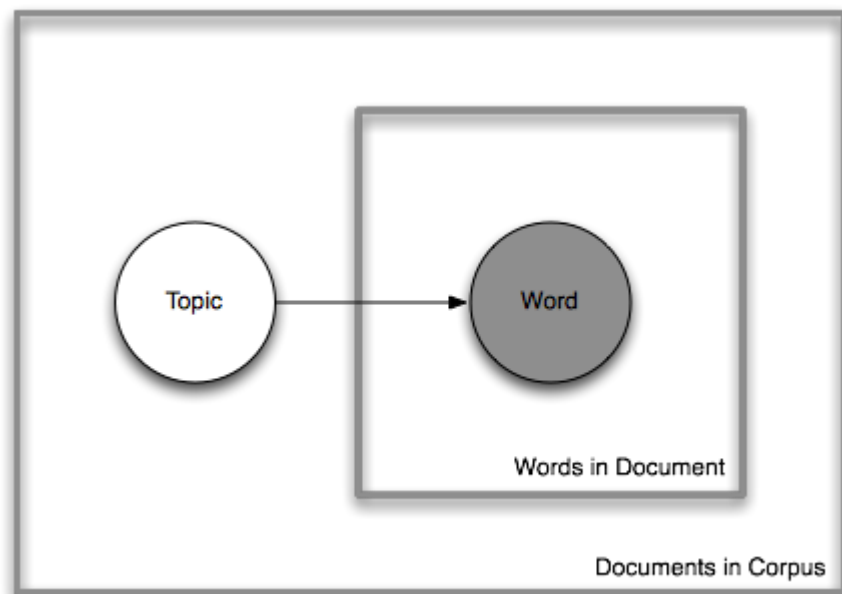
Dla każdego dokumentu w korpusie wykonaj następujące czynności:

Dla każdego słowa w dokumencie, wykonaj następujące czynności:

Wybierz słowo z rozkładu słów.

# Mieszanina unigramów

# Mieszanka unigramów



Dla każdego dokumentu w korpusie wykonaj następujące czynności:

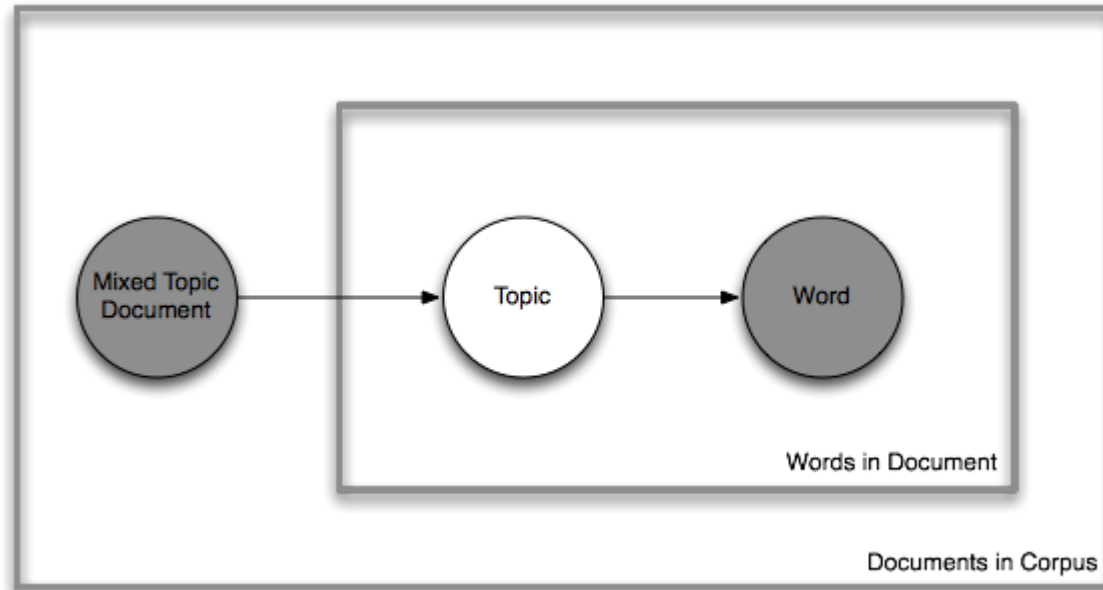
Wybierz rozkład słów (wątek) z rozkładu wątków.

Dla każdego słowa w dokumencie, wykonaj następujące czynności:

Wybierz słowo z tego rozkładu słów.

# Probabilistyczna analiza ukrytej semantyki

# Probabilistyczna analiza ukrytej semantyki



Dla każdego dokumentu w korpusie wykonaj następujące czynności:

Dla każdego słowa w dokumencie, wykonaj następujące czynności:

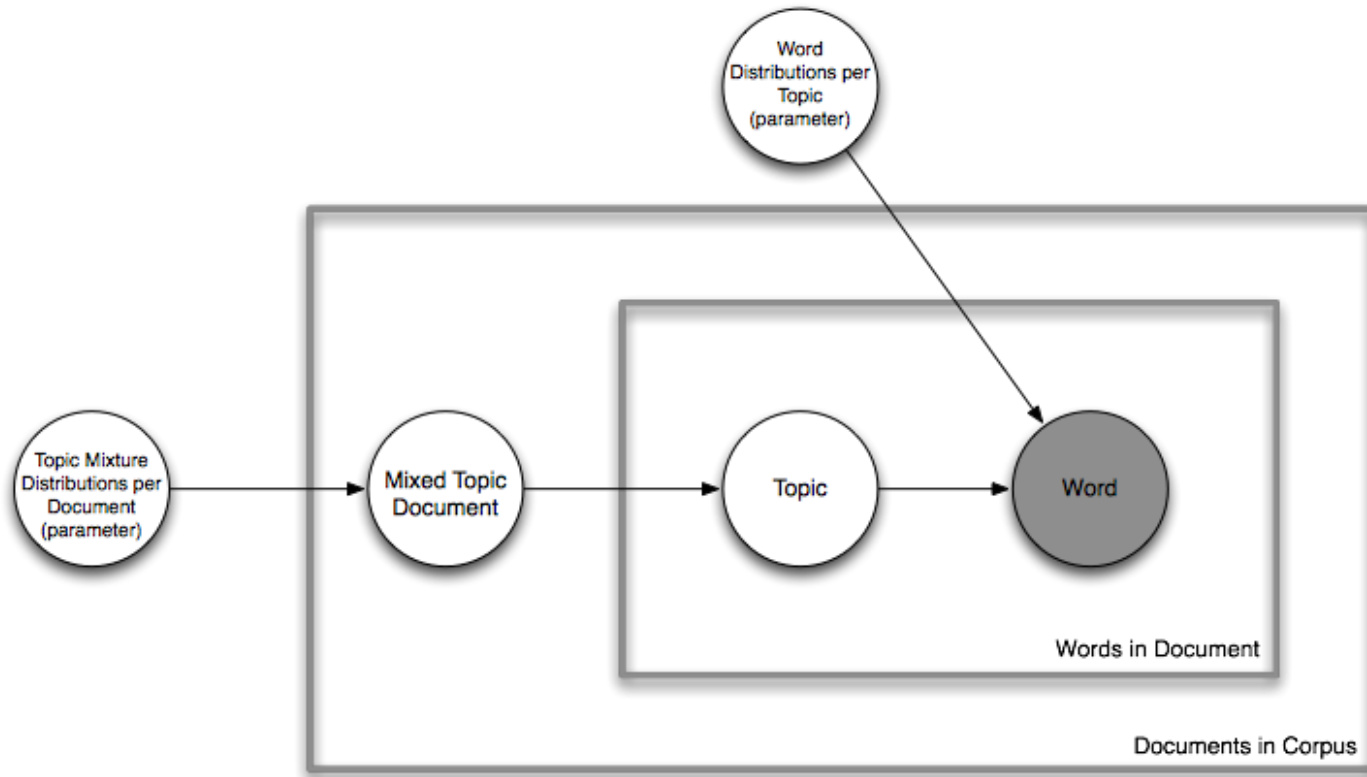
Wybierz rozkład słów z rozkładu wątków.

Wybierz słowo z tego rozkładu słów.

# Alokacja ukrytej zmiennej Dirichleta



# Alokacja ukrytej zmiennej Dirichleta



Dla każdego dokumentu w korpusie wykonać następujące kroki:

Wybierz rozkład mieszanki wątków z rozkładu Dirichleta.

Dla każdego słowa w dokumencie wykonaj następujące kroki:

Wybierz temat z rozkładu mieszanki wątków.

Wybierz słowo z rozkładu słów wybranego powyżej.

# LDA

- Model - traktuje dane jako obserwacje (wyrazy) procesu probabilistycznego (generowania dokumentu) z ukrytymi zmiennymi (tematami)

# LDA

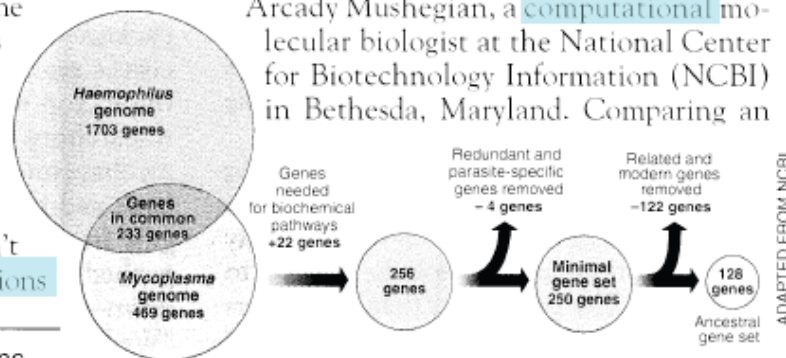
## Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

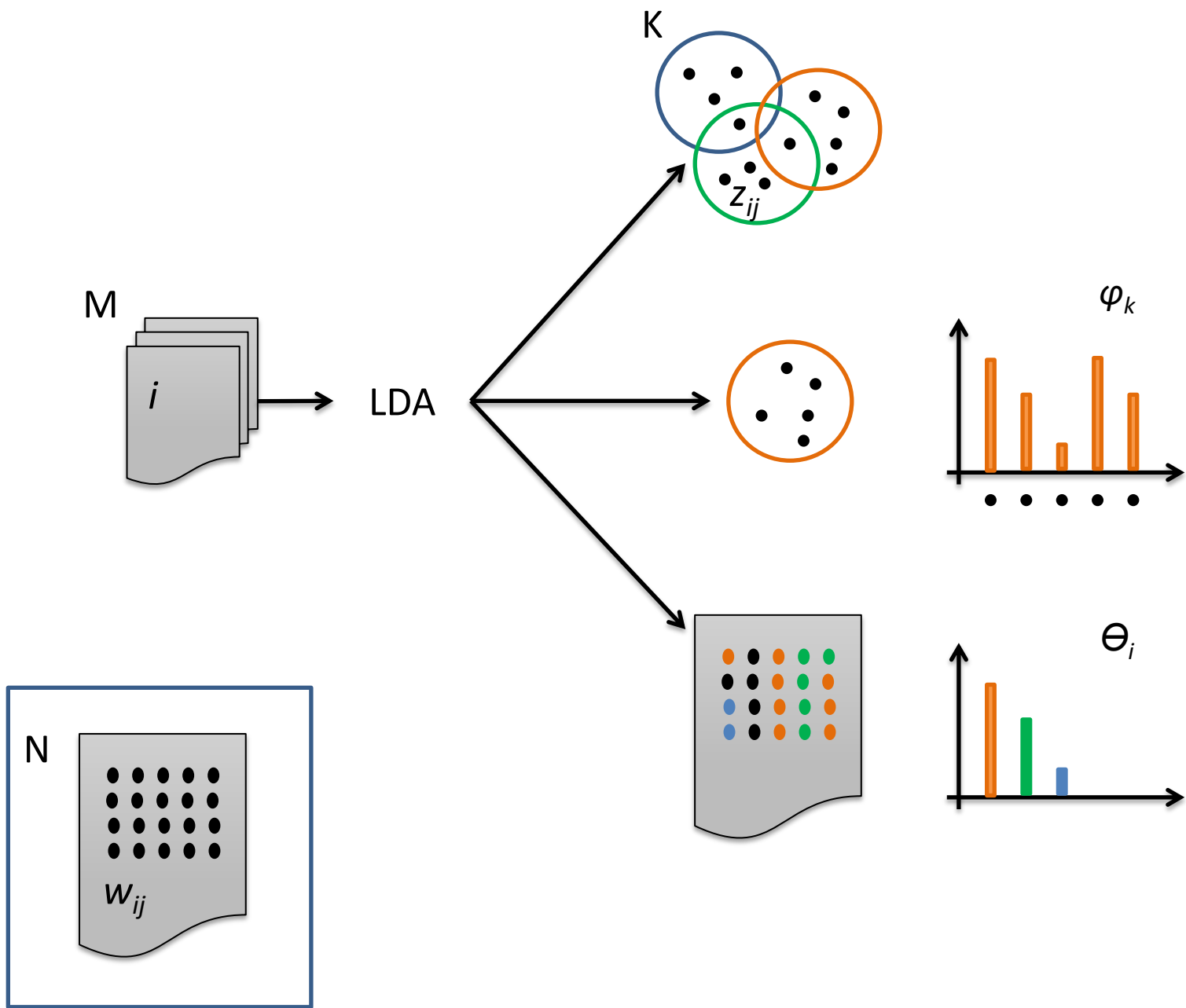
Although the numbers don't match precisely, those predictions

\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.



# LDA

## Topics

gene 0.04  
dna 0.02  
genetic 0.01  
...

life 0.02  
evolve 0.01  
organism 0.01  
...

brain 0.04  
neuron 0.02  
nerve 0.01  
...

data 0.02  
number 0.02  
computer 0.01  
...

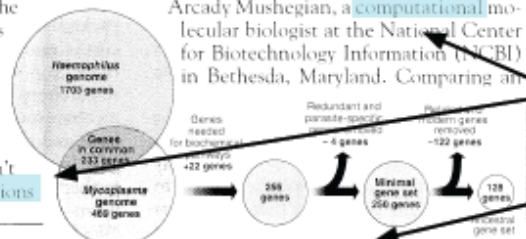
## Documents

### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson, a biologist at Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

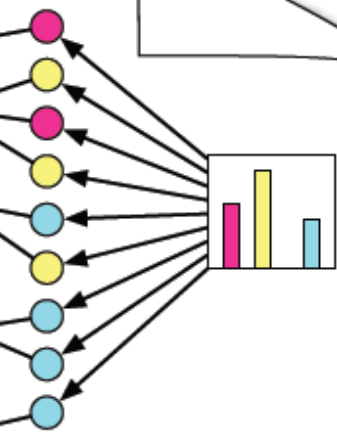


\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

## Topic proportions and assignments



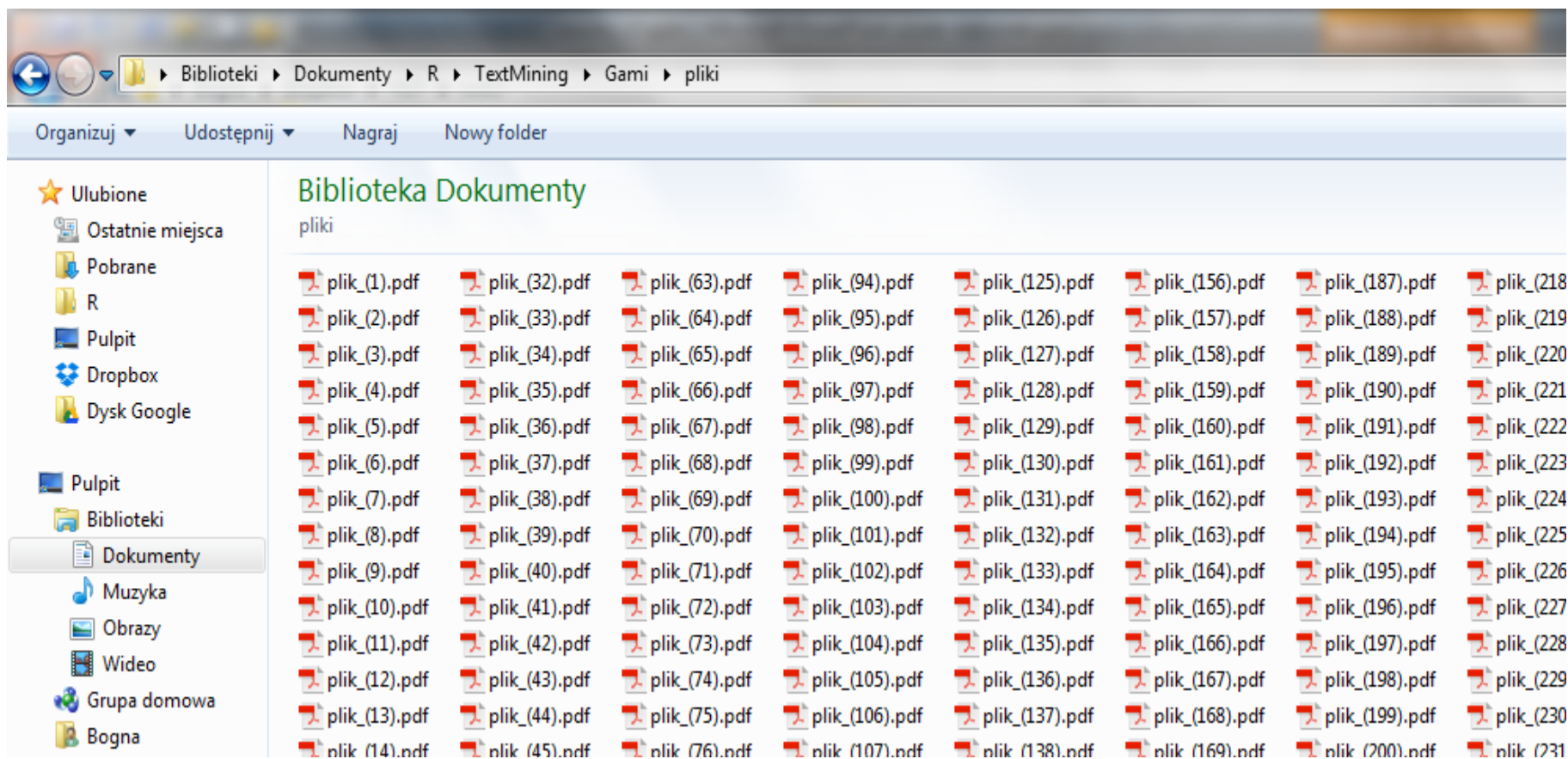
# Alokacja ukrytej zmiennej Dirichleta

- Ocena wyników:
  - Human-in-the-loop
    - Word intrusion
    - Topic intrusion
  - Metryki
    - Podobieństwo dokumentów

**PRZYKŁAD**

# Analiza wątków artykułów

- Analiza korpusu 444 artykułów (pdf)





# Analiza wątków artykułów

- Analiza korpusu 444 artykułów (pdf)
- Czyszczenie

```
removePunctuation = TRUE,  
stopwords = TRUE,  
tolower = TRUE,  
stemming = TRUE,  
removeNumbers = TRUE,  
bounds = list(global = c(3, Inf))
```

# Analiza wątków artykułów

- Analiza korpusu 444 artykułów (pdf)
- Czyszczenie
- Macierz Dokument-Term

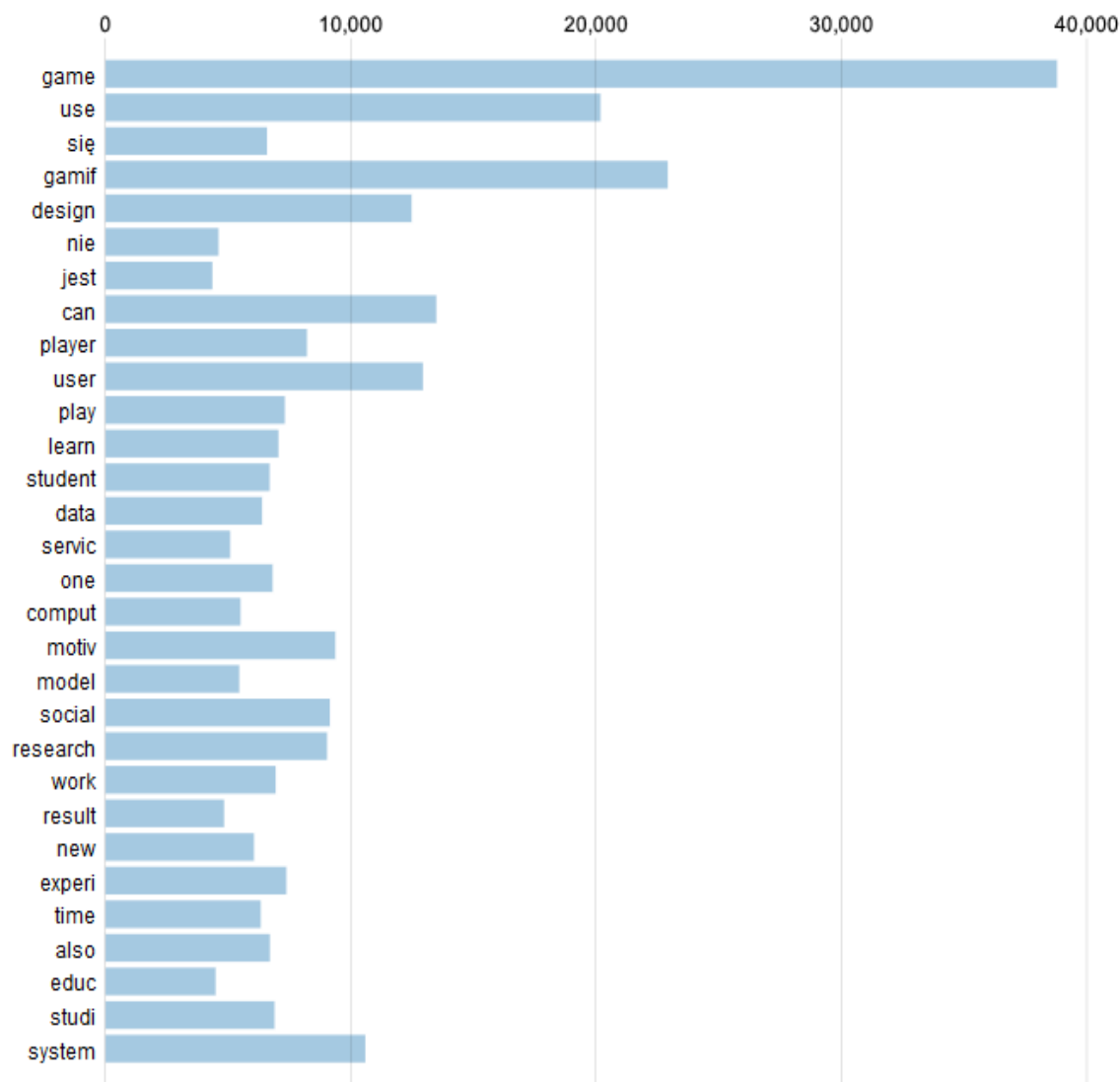
```
> inspect(Gami.tdm[1:10,1000:1100])
<<DocumentTermMatrix (documents: 10, terms: 101)>>
Non-/sparse entries: 125/885
Sparsity           : 88%
Maximal term length: 20
Weighting           : term frequency (tf)
Sample             :

```

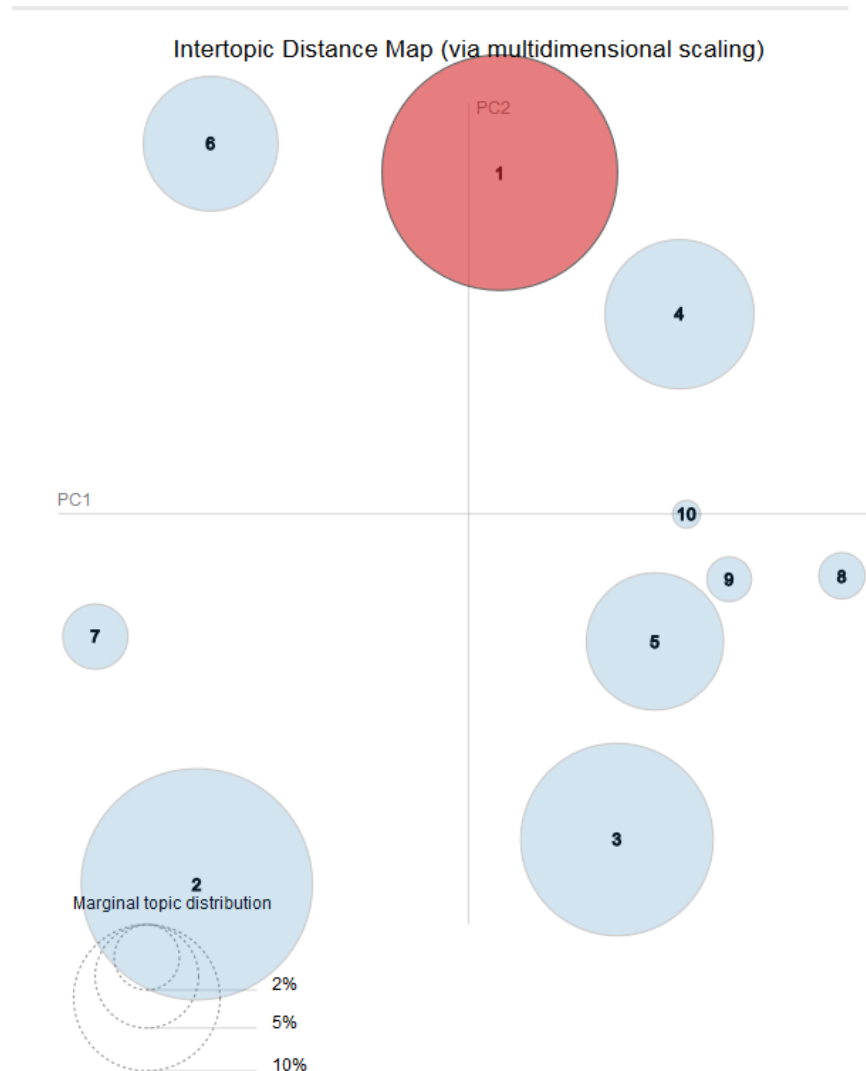
Docs	appeal	appear	appendix	appli	applic	appreci	approach	appropri	april	area
plik_(1).pdf	3	6	0	16	16	2	32	9	5	6
plik_(10).pdf	0	2	2	6	10	0	2	0	1	2
plik_(100).pdf	0	1	0	0	13	1	4	0	0	0
plik_(101).pdf	7	4	20	49	33	2	136	6	0	27
plik_(102).pdf	0	2	11	29	15	2	40	8	5	41
plik_(103).pdf	0	1	0	4	0	0	9	0	0	0
plik_(104).pdf	4	1	0	0	2	0	5	0	6	1
plik_(105).pdf	0	0	0	7	7	0	7	3	0	1
plik_(106).pdf	5	14	0	22	7	8	42	5	4	17
plik_(107).pdf	3	0	0	2	1	0	1	3	1	3

```
>
```

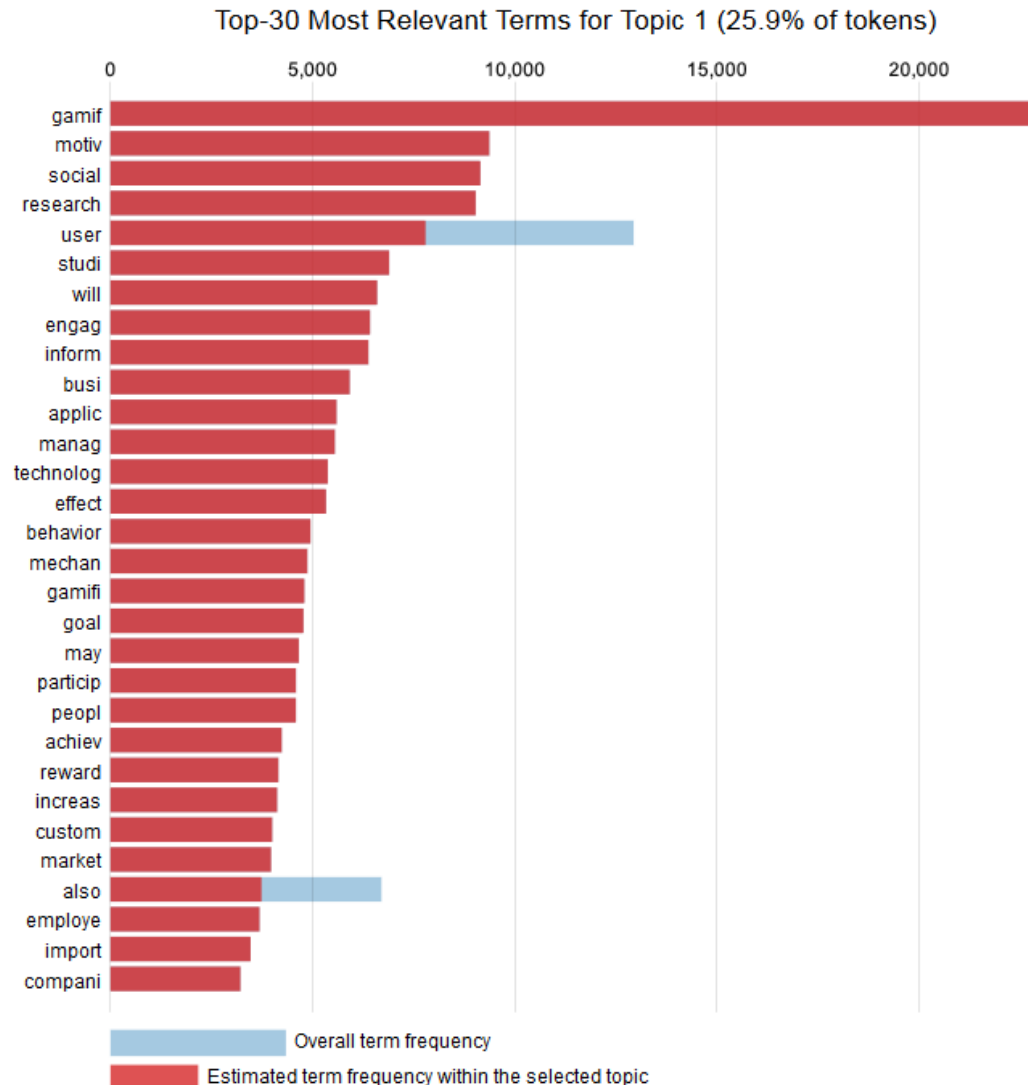
# Analiza wątków artykułów



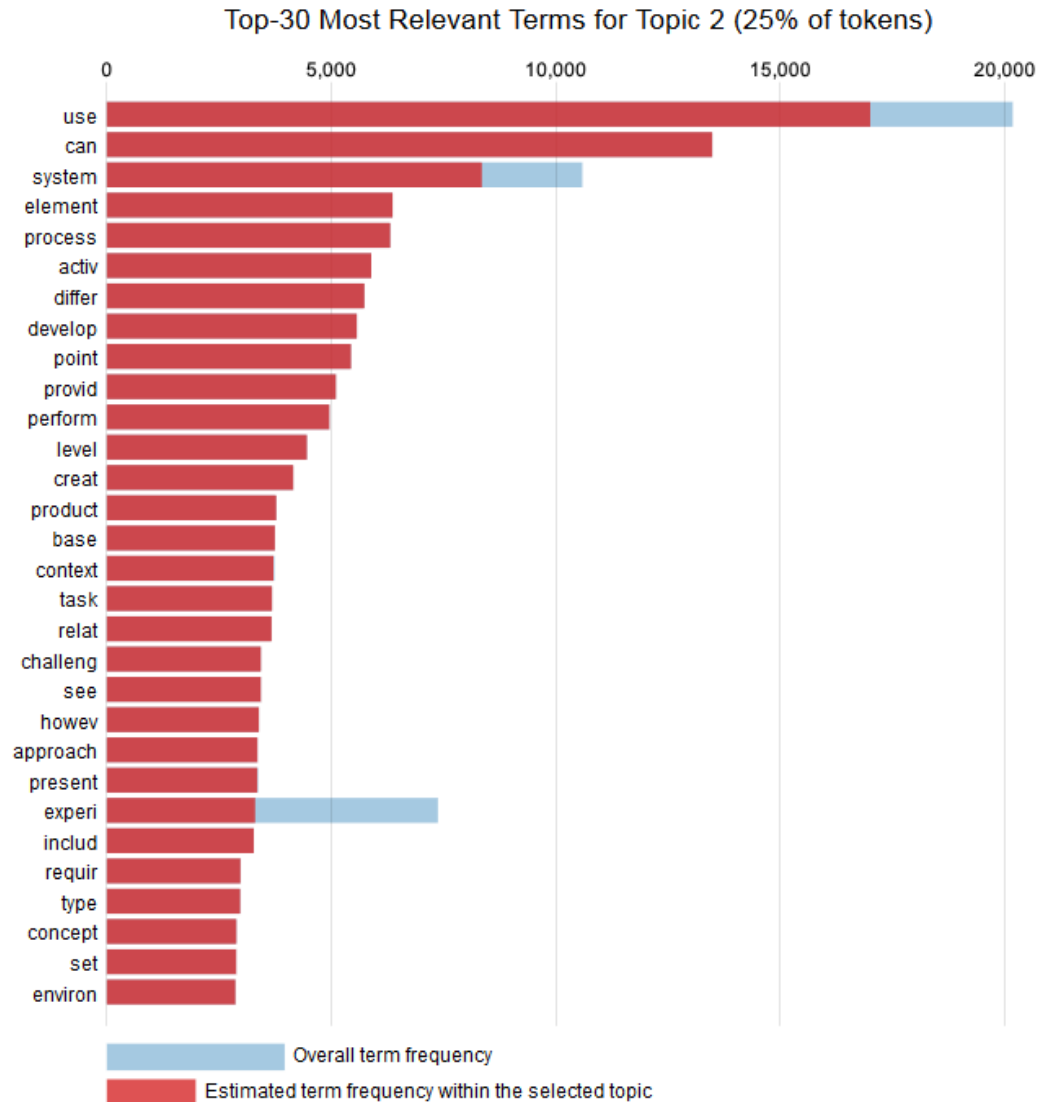
# Analiza wątków artykułów



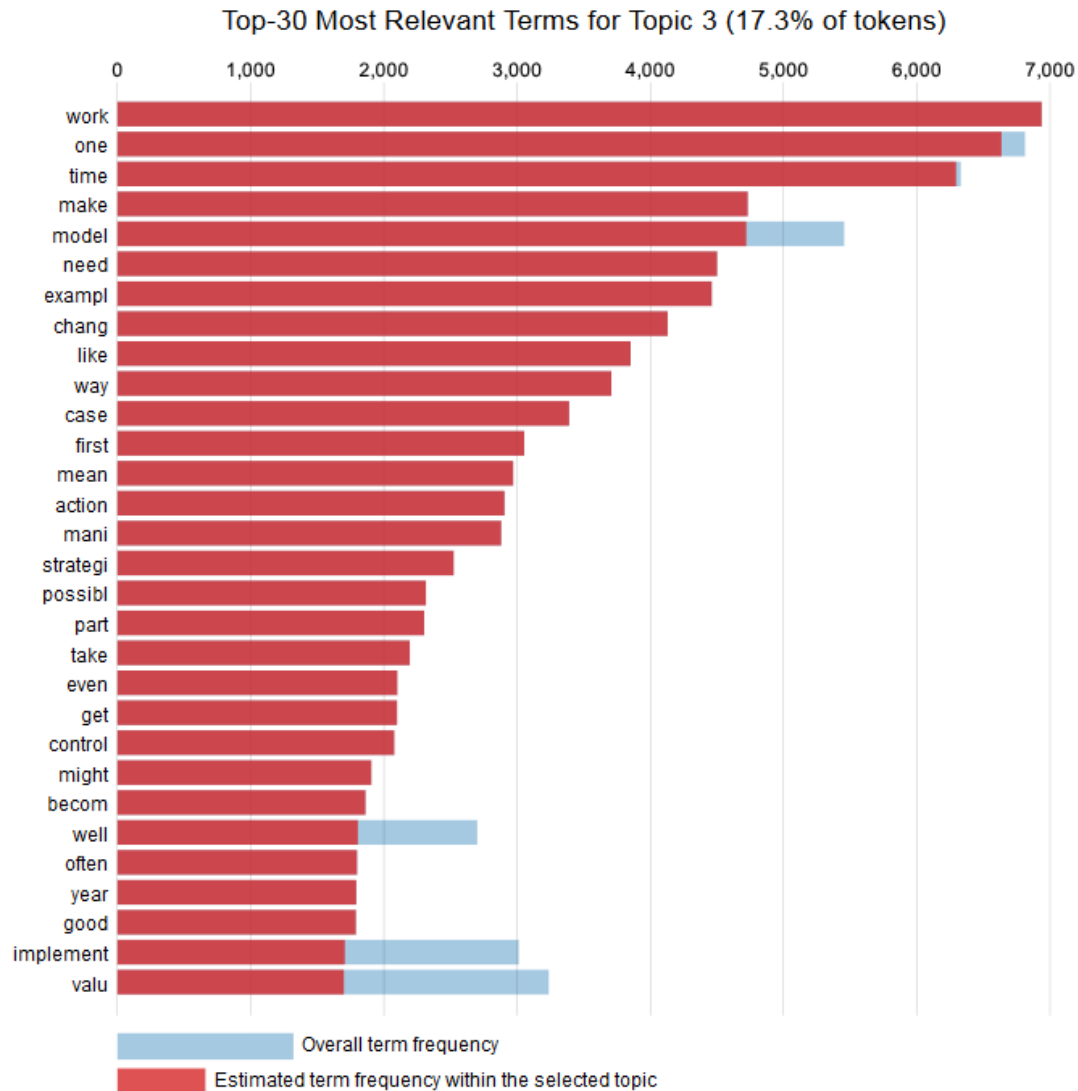
# Analiza wątków artykułów



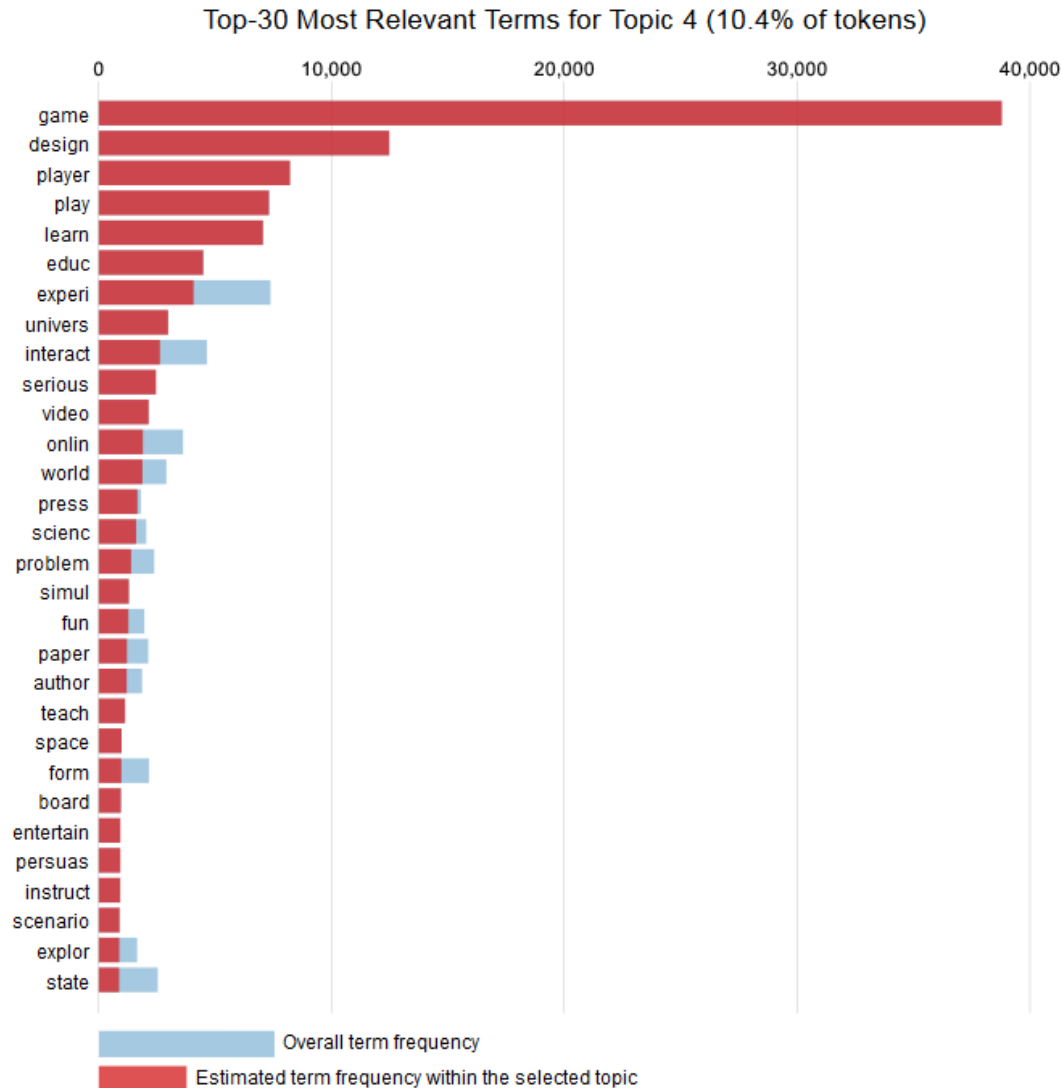
# Analiza wątków artykułów



# Analiza wątków artykułów



# Analiza wątków artykułów





# Literatura

- Kovanovic V., Topic Modeling for Learning Analytics Researchers LAK15 Tutorial  
[[www.slideshare.net/vitomirkovanovic/topic-modeling-for-learning-analytics-researchers-lak15-tutorial](http://www.slideshare.net/vitomirkovanovic/topic-modeling-for-learning-analytics-researchers-lak15-tutorial)]
- *Burton M.*, The Joy of Topic Modeling  
[[mcburton.net/blog/joy-of-tm/](http://mcburton.net/blog/joy-of-tm/)]